

# Lasso Regression

---

Nipun Batra

July 20, 2025

IIT Gandhinagar

- LASSO  $\rightarrow$  Least absolute shrinkage and selection operator

# Lasso Regression

- LASSO  $\rightarrow$  Least absolute shrinkage and selection operator
- Popular as it leads to a sparse solution.

# Constructing the Objective Function

- Find a  $\theta_{\text{opt}}$  such that

$$\theta_{\text{opt}} = \arg \min_{\theta} (\mathbf{y} - \mathbf{X}\theta)^T (\mathbf{y} - \mathbf{X}\theta) : \|\theta\|_1 < s \quad (1)$$

# Constructing the Objective Function

- Find a  $\theta_{\text{opt}}$  such that

$$\theta_{\text{opt}} = \arg \min_{\theta} (\mathbf{y} - \mathbf{X}\theta)^T (\mathbf{y} - \mathbf{X}\theta) : \|\theta\|_1 < s \quad (1)$$

- Using KKT conditions

$$\theta_{\text{opt}} = \arg \min_{\theta} \underbrace{(\mathbf{y} - \mathbf{X}\theta)^T (\mathbf{y} - \mathbf{X}\theta) + \delta^2 \|\theta\|_1}_{\text{convex function}} \quad (2)$$

## Solving the Objective

- Since  $\|\boldsymbol{\theta}\|_1$  is not differentiable, we cannot solve,

$$\frac{\partial(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \delta^2\|\boldsymbol{\theta}\|_1}{\partial\boldsymbol{\theta}} = 0 \quad (3)$$

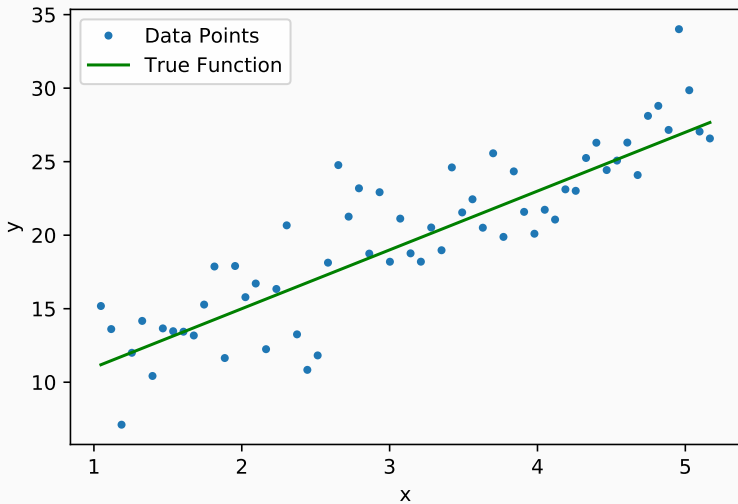
## Solving the Objective

- Since  $\|\boldsymbol{\theta}\|_1$  is not differentiable, we cannot solve,

$$\frac{\partial(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \delta^2\|\boldsymbol{\theta}\|_1}{\partial\boldsymbol{\theta}} = 0 \quad (3)$$

- How to Solve? Use coordinate descent!

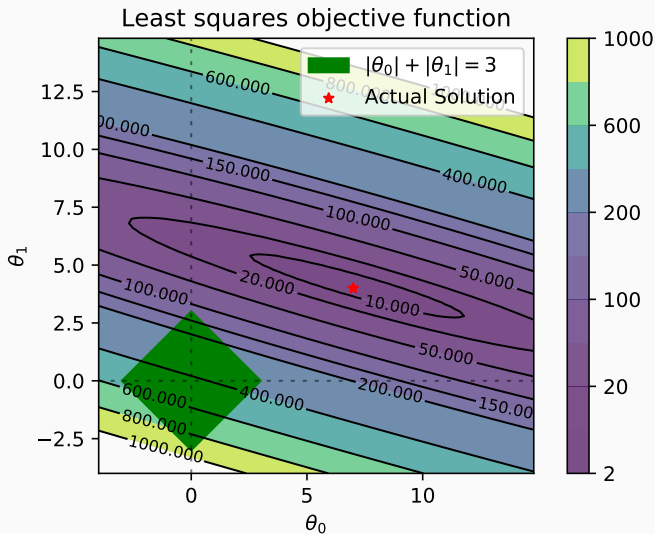
# Sample Dataset



$$y = 4x + 7$$

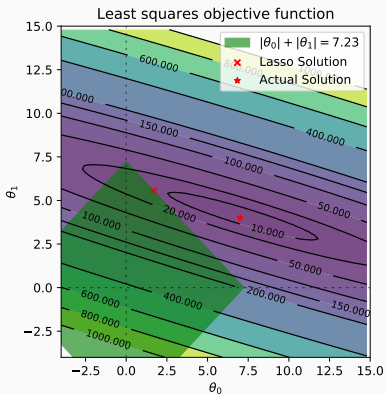
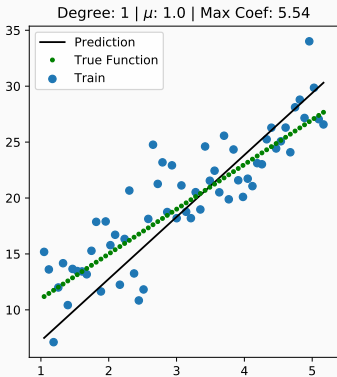


# Geometric Interpretation



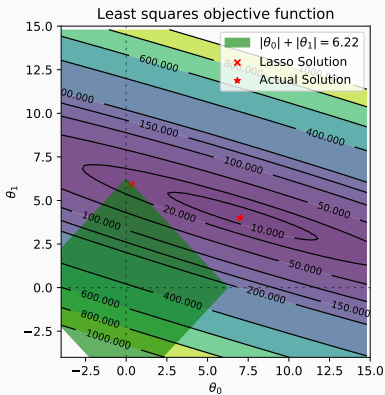
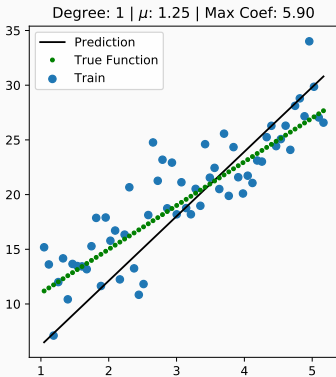
Lasso regression

# Effect of $\mu$ - Regularization of Parameters



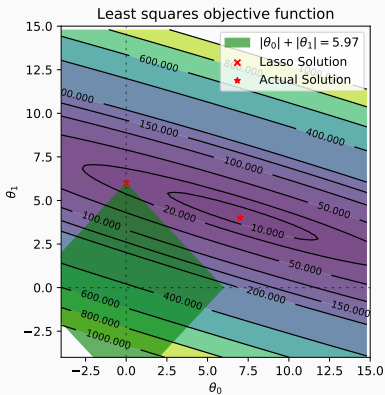
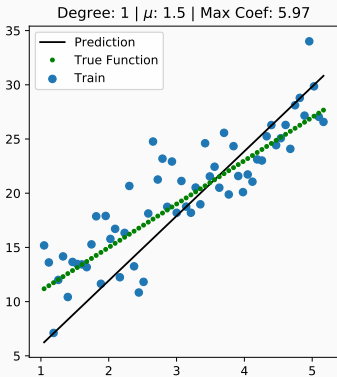
$\mu = 1.0$   
(on the *Sample Dataset*)

# Effect of $\mu$ - Regularization of Parameters



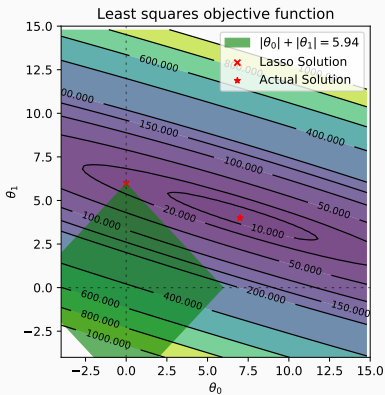
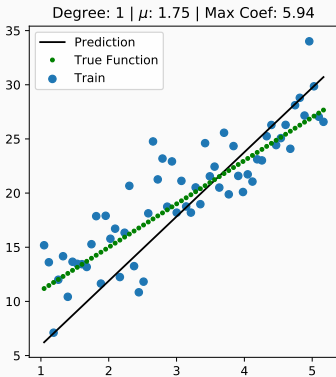
$\mu = 1.25$   
(on the *Sample Dataset*)

# Effect of $\mu$ - Regularization of Parameters



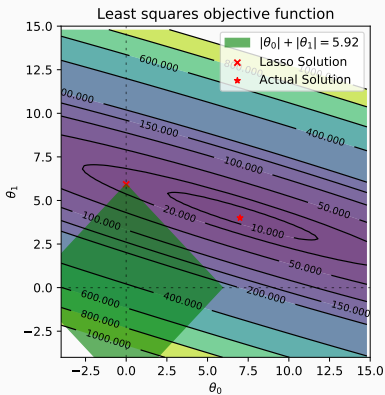
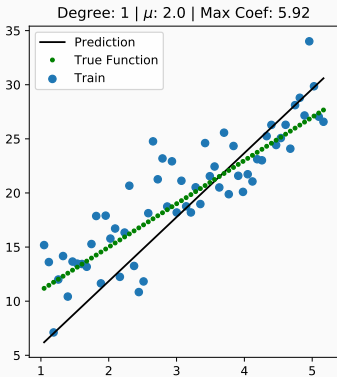
$\mu = 1.5$   
(on the *Sample Dataset*)

# Effect of $\mu$ - Regularization of Parameters



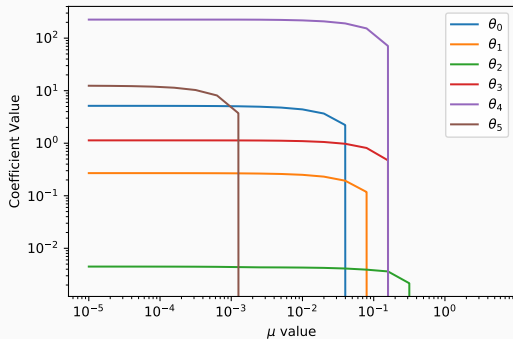
$\mu = 1.75$   
(on the *Sample Dataset*)

# Effect of $\mu$ - Regularization of Parameters



$\mu = 2.0$   
(on the *Sample Dataset*)

# Regularization path of lasso regression



Regularization path of  $\theta_i$

- LASSO inherently does feature selection!



# LASSO and feature selection

- LASSO inherently does feature selection!
- Sets coefficients of “less important” features to zero.

# LASSO and feature selection

- LASSO inherently does feature selection!
- Sets coefficients of “less important” features to zero.
- Sparse and memory efficient and often more interpretable models.

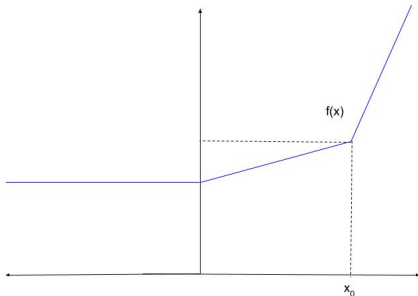
- Generalises gradient to convex but non-differentiable problems

- Generalises gradient to convex but non-differentiable problems
- Examples:

- Generalises gradient to convex but non-differentiable problems
- Examples:
  - $f(x) = |x|$

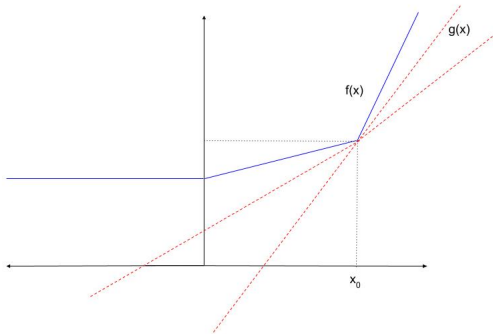
# Task at hand

- TASK: find derivative of  $f(x)$  at  $x = x_0$



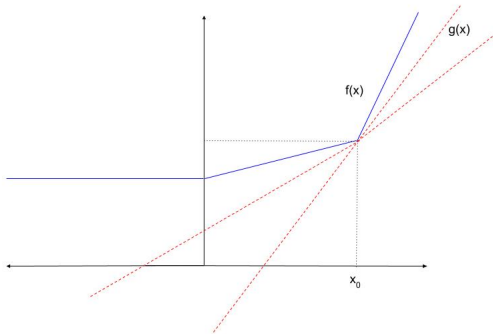
# Solution

- Construct a differentiable  $g(x)$



# Solution

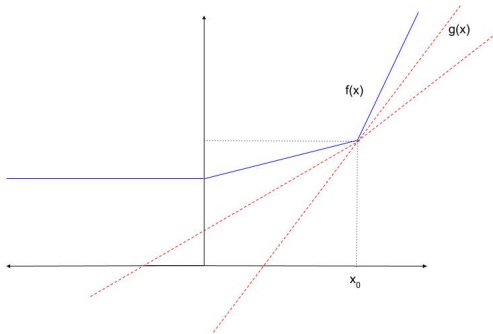
- Construct a differentiable  $g(x)$ 
  - Intersecting  $f(x)$  at  $x = x_0$





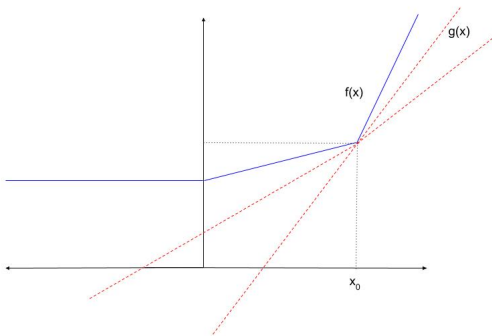
# Solution

- Construct a differentiable  $g(x)$ 
  - Intersecting  $f(x)$  at  $x = x_0$
  - Below or on  $f(x)$  for all  $x$



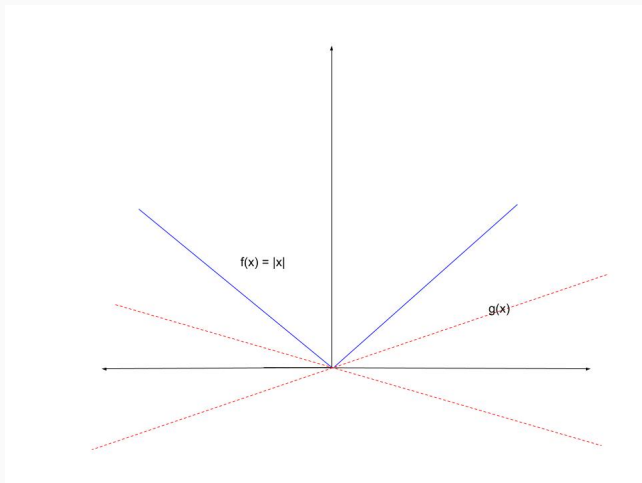
# Solution

- Compute slope of  $g(x)$  at  $x = x_0$



## Another Example: $f(x) = |x|$

- Subgradient of  $f(x)$  belongs to  $[-1, 1]$



- Another optimisation method (akin to gradient descent)

# Coordinate Descent

- Another optimisation method (akin to gradient descent)
- Objective:  $\min_{\theta} f(\theta)$

- Another optimisation method (akin to gradient descent)
- Objective:  $\min_{\theta} f(\theta)$
- Key idea: Sometimes difficult to find minimum for all coordinates

- Another optimisation method (akin to gradient descent)
- Objective:  $\min_{\theta} f(\theta)$
- Key idea: Sometimes difficult to find minimum for all coordinates
- ..., but, easy for each coordinate

# Coordinate Descent

- Another optimisation method (akin to gradient descent)
- Objective:  $\min_{\theta} f(\theta)$
- Key idea: Sometimes difficult to find minimum for all coordinates
- ..., but, easy for each coordinate
- turns into a one-dimensional optimisation problem



- Picking next coordinate:

- Picking next coordinate:

- Picking next coordinate: random, round-robin
- No step-size to choose!

# Coordinate Descent

- Picking next coordinate: random, round-robin
- No step-size to choose!
- Converges for Lasso objective

## Coordinate Descent : Example

Learn  $y = \theta_0 + \theta_1 x$  on following dataset, using coordinate descent where initially  $(\theta_0, \theta_1) = (2, 3)$  for 2 iterations.

<b>x</b>	<b>y</b>
1	1
2	2
3	3

## Coordinate Descent : Example

Our predictor,  $\hat{y} = \theta_0 + \theta_1 x$

Error for  $i^{th}$  datapoint,  $\epsilon_i = y_i - \hat{y}_i$

$$\epsilon_1 = 1 - \theta_0 - \theta_1$$

$$\epsilon_2 = 2 - \theta_0 - 2\theta_1$$

$$\epsilon_3 = 3 - \theta_0 - 3\theta_1$$

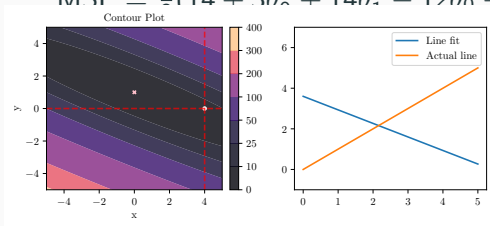
$$\text{MSE} = \frac{\epsilon_1^2 + \epsilon_2^2 + \epsilon_3^2}{3} = \frac{14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1}{3}$$

## Iteration 0

$$\text{MSE} = \frac{1}{3}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$$

# Iteration 0

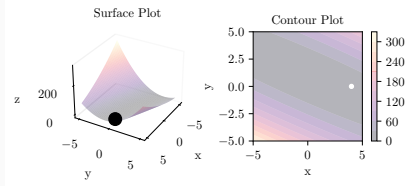
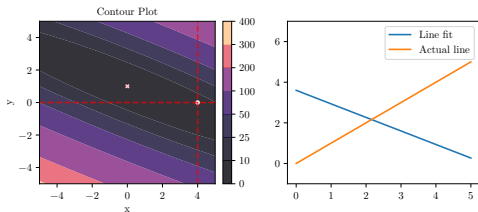
$$MSF = \frac{1}{2}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$$





# Iteration 0

$$MSF = \frac{1}{2}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$$



# Coordinate Descent : Example

## Iteration 1

INIT:  $\theta_0 = 2$  and  $\theta_1 = 3$

$\theta_1 = 3$  optimize for  $\theta_0$

# Coordinate Descent : Example

## Iteration 1

INIT:  $\theta_0 = 2$  and  $\theta_1 = 3$

$\theta_1 = 3$  optimize for  $\theta_0$

$$\frac{\partial \text{MSE}}{\partial \theta_0} = 6\theta_0 + 24 = 0$$

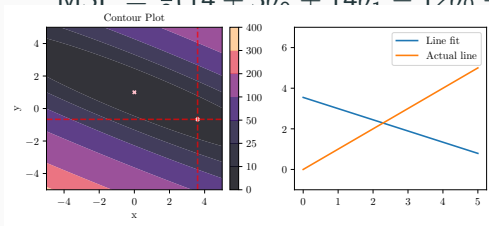
$$\theta_0 = -4$$

## Iteration 1

$$\text{MSE} = \frac{1}{3}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$$

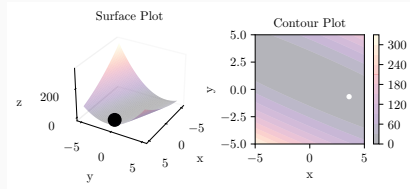
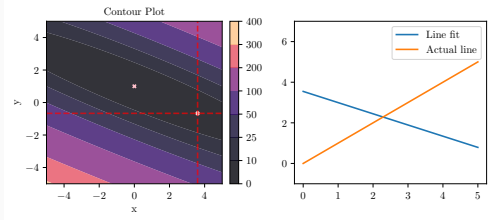
# Iteration 1

$$MSF = \frac{1}{2}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$$



# Iteration 1

$$MSF = \frac{1}{2}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$$



# Coordinate Descent : Example

## Iteration 2

INIT:  $\theta_0 = -4$  and  $\theta_1 = 3$

$\theta_0 = -4$  optimize for  $\theta_1$

# Coordinate Descent : Example

## Iteration 2

INIT:  $\theta_0 = -4$  and  $\theta_1 = 3$

$\theta_0 = -4$  optimize for  $\theta_1$

$\theta_1 = 2.7$

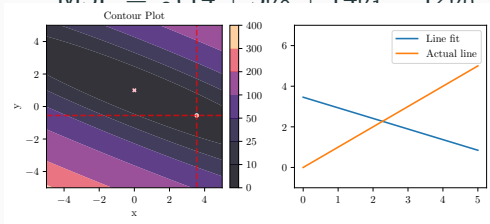


## Iteration 2

$$\text{MSE} = \frac{1}{3}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$$

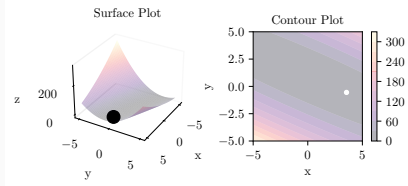
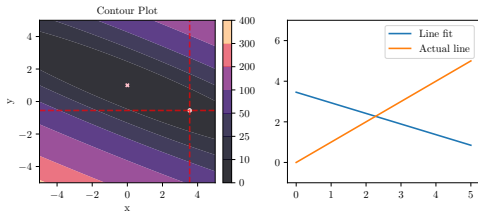
## Iteration 2

$$MSF = \frac{1}{2}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$$



## Iteration 2

$$MSF = \frac{1}{2}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$$



# Coordinate Descent : Example

## Iteration 3

INIT:  $\theta_0 = -4$  and  $\theta_1 = 2.7$

$\theta_1 = 2.7$  optimize for  $\theta_0$

# Coordinate Descent : Example

## Iteration 3

INIT:  $\theta_0 = -4$  and  $\theta_1 = 2.7$

$\theta_1 = 2.7$  optimize for  $\theta_0$

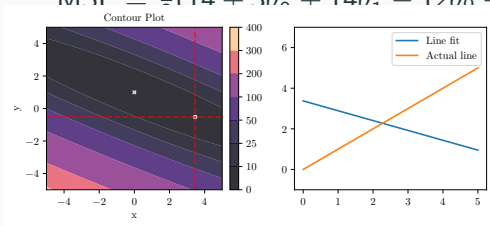
$\theta_0 = -3.4$

## Iteration 3

$$\text{MSE} = \frac{1}{3}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$$

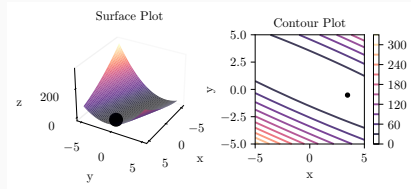
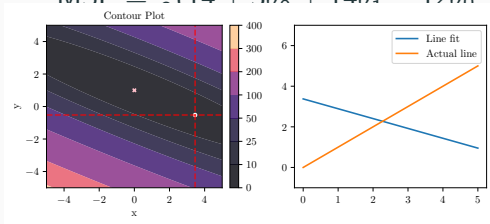
# Iteration 3

$$MSF = \frac{1}{2}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$$



# Iteration 3

$$MSF = \frac{1}{2}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$$





# Coordinate Descent for Unregularised Regression

- Express error as a difference of  $y_i$  and  $\hat{y}_i$

$$\hat{y}_i = \sum_{j=0}^d \theta_j x_i^j = \theta_0 x_i^0 + \theta_1 x_i^1 + \theta_2 x_i^2 + \dots + \theta_d x_i^d \quad (4)$$

$$\epsilon_i = y_i - \hat{y}_i = y_i - \theta_0 x_i^0 - \theta_1 x_i^1 - \dots - \theta_d x_i^d = y_i - \sum_{j=0}^d \theta_j x_i^j \quad (5)$$

## Coordinate Descent for Unregularised regression

$$\sum_{i=1}^n \epsilon^2 == \sum_{i=1}^n \left( y_i - \left( \theta_0 x_i^0 + \dots + \theta_j x_i^j + \theta_d x_i^d \right) \right)^2$$

## Coordinate Descent for Unregularised regression

$$\sum_{i=1}^n \epsilon^2 == \sum_{i=1}^n \left( y_i - \left( \theta_0 x_i^0 + \dots + \theta_j x_i^j + \theta_d x_i^d \right) \right)^2$$

$$\frac{\partial (\theta_j)}{\partial \theta_j} = 2 \sum_{i=1}^n \left( y_i - \left( \theta_0 x_i^0 + \dots + \theta_j x_i^j + \dots \right) \right) \left( -x_i^j \right)$$

## Coordinate Descent for Unregularised regression

$$\begin{aligned}\sum_{i=1}^n \epsilon^2 &= \sum_{i=1}^n \left( y_i - \left( \theta_0 x_i^0 + \dots + \theta_j x_i^j + \theta_d x_i^d \right) \right)^2 \\ \frac{\partial (\theta_j)}{\partial \theta_j} &= 2 \sum_{i=1}^n \left( y_i - \left( \theta_0 x_i^0 + \dots + \theta_j x_i^j + \dots \right) \right) \left( -x_i^j \right) \\ &= 2 \sum_{i=1}^n \left( y_i - \left( \theta_0 x_i^0 + \dots + \theta_d x_i^d \right) \right) \left( -x_i^j \right) + 2 \sum_{i=1}^n \theta_j (x_i^j)^2\end{aligned}$$

## Coordinate Descent for Unregularised regression

$$\begin{aligned}\sum_{i=1}^n \epsilon^2 &= \sum_{i=1}^n \left( y_i - \left( \theta_0 x_i^0 + \dots + \theta_j x_i^j + \theta_d x_i^d \right) \right)^2 \\ \frac{\partial (\theta_j)}{\partial \theta_j} &= 2 \sum_{i=1}^n \left( y_i - \left( \theta_0 x_i^0 + \dots + \theta_j x_i^j + \dots \right) \right) \left( -x_i^j \right) \\ &= 2 \sum_{i=1}^n \left( y_i - \left( \theta_0 x_i^0 + \dots + \theta_d x_i^d \right) \right) \left( -x_i^j \right) + 2 \sum_{i=1}^n \theta_j (x_i^j)^2\end{aligned}$$

where:

$$\hat{y}_i^{(-j)} = \theta_0 x_i^0 + \dots + \theta_d x_i^d$$

is  $\hat{y}_i$  without  $\theta_j$

## Coordinate Descent for Unregularised regression

$$\text{Set } \frac{\partial (\theta_j)}{\partial \theta_j} = 0$$

$$\theta_j = \sum_{i=1}^n \frac{(y_i - (\theta_0 x_i^0 + \dots + \theta_d x_i^d)) (x_i^j)}{(x_i^j)^2} = \frac{\rho_j}{z_j}$$

$$\rho_j = \sum_{i=1}^n x_i^j (y_i - \hat{y}_i^{(-j)}) \quad \text{and} \quad z_j = \sum_{i=1}^n (x_i^j)^2$$

$z_j$  is the squared of  $\ell_2$  norm of the  $j^{th}$  feature

# Coordinate Descent for Lasso Regression

$$\text{Minimise } \underbrace{\sum_{i=1}^n \epsilon^2 + \delta^2 \{|\theta_0| + |\theta_1| + \dots |\theta_j| + \dots |\theta_d|\}}_{\text{LASSO OBJECTIVE}}$$

$$\frac{\partial}{\partial \theta_j} (\text{LASSO OBJECTIVE}) = -2\rho_j + 2\theta_j z_j + \delta^2 \frac{\partial}{\partial \theta_j} |\theta_j|$$

$$\frac{\partial}{\partial \theta_j} |\theta_j| = \begin{cases} 1 & \theta_j > 0 \\ [-1, 1] & \theta_j = 0 \\ -1 & \theta_j < 0 \end{cases}$$

# Coordinate Descent for Lasso Regression

- **Case 1:**  $\theta_j > 0$

$$-2\rho_j + 2\theta_j z_j + \delta^2 = 0$$

$$\theta_j = \frac{\rho_j - \frac{\delta^2}{2}}{z_j}$$

$$\rho_j > \frac{\delta^2}{2} \Rightarrow \theta_j = \frac{\rho_j - \frac{\delta^2}{2}}{z_j}$$



# Coordinate Descent for Lasso Regression

- **Case 1:**  $\theta_j > 0$

$$-2\rho_j + 2\theta_j z_j + \delta^2 = 0$$

$$\theta_j = \frac{\rho_j - \frac{\delta^2}{2}}{z_j}$$

$$\rho_j > \frac{\delta^2}{2} \Rightarrow \theta_j = \frac{\rho_j - \frac{\delta^2}{2}}{z_j}$$

- **Case 2:**  $\theta_j < 0$

$$\rho_j < \frac{\delta^2}{2} \Rightarrow \theta_j = \frac{\rho_j + \delta^2/2}{z_j} \tag{6}$$

# Coordinate Descent for Lasso Regression

- **Case 3:**  $\theta_j = 0$

$$\frac{\partial}{\partial \theta_j}(\text{LASSO OBJECTIVE}) = -2\rho_j + 2\theta_j z_j + \underbrace{\delta^2 \frac{\partial}{\partial \theta_j} |\theta_j|}_{[-1,1]}$$

$$\in \underbrace{[-2\rho_j - \delta^2, -2\rho_j + \delta^2]}_{\{0\} \text{ lies in this range}}$$

$$-2\rho_j - \delta^2 \leq 0 \text{ and } -2\rho_j + \delta^2 \geq 0$$

$$-\frac{\delta^2}{2} \leq \rho_j \leq \frac{\delta^2}{2} \Rightarrow \theta_j = 0$$

## Summary of Lasso Regression

$$\theta_j = \begin{bmatrix} \frac{\rho_j + \frac{\delta^2}{2}}{z_j} & \text{if } \rho_j < -\frac{\delta^2}{2} \\ 0 & \text{if } -\frac{\delta^2}{2} \leq \rho_j \leq \frac{\delta^2}{2} \\ \frac{\rho_j - \frac{\delta^2}{2}}{z_j} & \text{if } \rho_j > \frac{\delta^2}{2} \end{bmatrix} \quad (7)$$